Docket No.: 67161-141 **PATENT**

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of	:	Customer Number: 20277
Ryuji MANO	:	Confirmation Number:
Serial No.:	:	Group Art Unit:
Filed: February 12, 2004	:	Examiner:

For: SPEECH RECOGNITION APPARATUS CAPABLE OF IMPROVING RECOGNITION RATE REGARDLESS OF AVERAGE DURATION OF PHONEMES

CLAIM OF PRIORITY AND
TRANSMITTAL OF CERTIFIED PRIORITY DOCUMENT

Mail Stop CPD
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

In accordance with the provisions of 35 U.S.C. 119, Applicant hereby claim the priority of:

Japanese Patent Application No. JP 2003-277661, filed on July 22, 2003.

cited in the Declaration of the present application. A certified copy is submitted herewith.

Respectfully submitted,

MCDERMOTT, WILL & EMERY

Stephen A. Becker
Registration No. 26,527

600 13th Street, N.W.
Washington, DC 20005-3096
(202) 756-8000 SAB:gav
Facsimile: (202) 756-8087
Date: February 12, 2004

67161-141
Ryuji MANO
February 12, 2004

日本国特許庁 JAPAN PATENT OFFICE

McDermott, Will & Emery

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: ２００３年　７月２２日

出願番号
Application Number: 特願２００３－２７７６６１

[ST. 10/C]: [ＪＰ２００３－２７７６６１]

出願人
Applicant(s): 株式会社ルネサステクノロジ

２００３年　８月２６日

特許庁長官
Commissioner,
Japan Patent Office 今井康夫

出証番号　出証特２００３－３０６９６６１

【書類名】　特許願
【整理番号】　544000JP01
【提出日】　平成15年 7月22日
【あて先】　特許庁長官殿
【国際特許分類】　G10L 3/00
G10L 15/02
【発明者】
【住所又は居所】　東京都千代田区丸の内二丁目４番１号　株式会社ルネサステクノロジ内
【氏名】　眞野　竜二
【特許出願人】
【識別番号】　503121103
【氏名又は名称】　株式会社ルネサステクノロジ
【代理人】
【識別番号】　100064746
【弁理士】
【氏名又は名称】　深見　久郎
【選任した代理人】
【識別番号】　100085132
【弁理士】
【氏名又は名称】　森田　俊雄
【選任した代理人】
【識別番号】　100083703
【弁理士】
【氏名又は名称】　仲村　義平
【選任した代理人】
【識別番号】　100096781
【弁理士】
【氏名又は名称】　堀井　豊
【選任した代理人】
【識別番号】　100098316
【弁理士】
【氏名又は名称】　野田　久登
【選任した代理人】
【識別番号】　100109162
【弁理士】
【氏名又は名称】　酒井　將行
【手数料の表示】
【予納台帳番号】　008693
【納付金額】　21,000円
【提出物件の目録】
【物件名】　特許請求の範囲　1
【物件名】　明細書　1
【物件名】　図面　1
【物件名】　要約書　1

【書類名】特許請求の範囲
【請求項１】
入力音声信号に対して、各々が所定時間長を有する時間窓に対応する複数のフレームを、少なくとも異なる時間幅でスライドさせることにより、特徴パラメータの抽出を行う特徴抽出手段と、
前記入力音声の音韻パターンにそれぞれ対応づけて標準パターンデータを格納するための記憶手段と、
前記特徴抽出手段で抽出された前記特徴パラメータと前記標準パターンデータを照合することで、対応する音韻を認識し、認識結果を出力するための認識手段とを備える、音声認識装置。
【請求項２】
前記特徴抽出手段は、前記入力音声信号の語頭から語尾に渡って、前記フレームをスライドさせる時間幅を順次増加させ、
前記記憶手段は、前記特徴抽出手段が前記フレームをスライドさせる時間幅のパターンに対応する前記標準パターンデータを予め格納する、請求項１記載の音声認識装置。
【請求項３】
前記特徴抽出手段は、
第１の固定時間幅で前記フレームをスライドさせつつ、前記特徴パラメータを抽出する第１の固定フレーム間隔抽出処理手段と、
前記第１の固定時間幅よりも短い第２の固定時間幅で前記時間窓をスライドさせつつ、前記特徴パラメータを抽出する第２の固定フレーム間隔抽出処理手段とを含み、
前記標準パターンデータは、前記第１の固定フレーム間隔抽出処理手段が前記フレームをスライドさせる時間幅の第１のパターンに対応する第１の標準パターンデータと、前記第２の固定フレーム間隔抽出処理手段が前記フレームをスライドさせる時間幅の第２のパターンに対応する第２の標準パターンデータとを含む、請求項１記載の音声認識装置。
【請求項４】
前記特徴抽出手段は、
固定時間幅で前記フレームをスライドさせつつ、前記特徴パラメータを抽出する固定フレーム間隔抽出処理手段と、
前記入力音声信号の語頭から語尾に渡って、前記フレームをスライドさせる時間幅を順次増加させつつ、前記特徴パラメータを抽出する可変フレーム間隔抽出処理手段とを含み、
前記標準パターンデータは、前記固定フレーム間隔抽出処理手段が前記フレームをスライドさせる時間幅の第１のパターンに対応する第１の標準パターンデータと、前記可変フレーム間隔抽出処理手段が前記フレームをスライドさせる時間幅の第２のパターンに対応する第２の標準パターンデータとを含む、請求項１記載の音声認識装置。
【請求項５】
前記特徴抽出手段は、
第１の固定時間幅で前記フレームをスライドさせつつ、前記特徴パラメータを抽出する第１の固定フレーム間隔抽出処理手段と、
前記第１の固定時間幅よりも短い第２の固定時間幅で前記時間窓をスライドさせつつ、前記特徴パラメータを抽出する第２の固定フレーム間隔抽出処理手段とを含み、
前記標準パターンデータは、前記第１の固定フレーム間隔抽出処理手段が前記フレームをスライドさせる時間幅の第１のパターンに対応する第１の標準パターンデータと、前記第２の固定フレーム間隔抽出処理手段が前記フレームをスライドさせる時間幅の第２のパターンに対応する第２の標準パターンデータとを含み、
前記入力音声信号と前記特徴抽出手段との間に設けられ、前記第１の固定フレーム間隔抽出処理手段から抽出された前記特徴パラメータに基づく、前記認識手段による照合結果に応じて、前記入力音声信号を前記第１の固定フレーム間隔抽出処理手段から前記第２の固定フレーム間隔抽出処理手段に切替えて与える入力選択手段をさらに備える、請求項１

記載の音声認識装置。
【請求項６】
前記第１の標準パターンデータは時刻と関連付けられており、
前記第１の標準パターンデータに基づいて、前記第２の標準パターンデータを補間により生成する補間処理手段をさらに備える、請求項５記載の音声認識装置。
【請求項７】
前記第１の標準パターンデータおよび第２の標準パターンデータは時刻と関連付けられており、
前記第２の固定フレーム間隔抽出処理手段が前記フレームをスライドさせる各時刻点は、前記第１の固定フレーム間隔抽出処理手段が前記フレームをスライドさせる時刻点のいずれかに対応する、請求項６記載の音声認識装置。
【請求項８】
前記特徴抽出手段は、
固定時間幅で前記フレームをスライドさせつつ、前記特徴パラメータを抽出する固定フレーム間隔抽出処理手段と、
前記入力音声信号の語頭から語尾に渡って、前記フレームをスライドさせる時間幅を順次増加させつつ、前記特徴パラメータを抽出する可変フレーム間隔抽出処理手段とを含み、
前記標準パターンデータは、前記固定フレーム間隔抽出処理手段が前記フレームをスライドさせる時間幅の第１のパターンに対応する第１の標準パターンデータと、前記可変フレーム間隔抽出処理手段が前記フレームをスライドさせる時間幅の第２のパターンに対応する第２の標準パターンデータとを含み、
前記入力音声信号と前記特徴抽出手段との間に設けられ、前記固定フレーム間隔抽出処理手段から抽出された前記特徴パラメータに基づく、前記認識手段による照合結果に応じて、前記入力音声信号を前記固定フレーム間隔抽出処理手段から前記可変フレーム間隔抽出処理手段に切替えて与える入力選択手段をさらに備える、請求項１記載の音声認識装置。
【請求項９】
前記第１の標準パターンデータは時刻と関連付けられており、
前記第１の標準パターンデータに基づいて、前記第２の標準パターンデータを補間により生成する補間処理手段をさらに備える、請求項８記載の音声認識装置。

【書類名】明細書
【発明の名称】音声認識装置
【技術分野】
【０００１】
本発明は、音韻単位の認識に基づく音声認織装置の構成に関するものである。
【背景技術】
【０００２】
従来、音声認識装置における音声認識は、そのほとんどが音声を特徴量の時系列に変換し、その時系列をあらかじめもっている標準パターンの時系列と比較して認識を行うことにより実行されている。
【０００３】
たとえば、特開２００１－３５６７９０号公報では、人の音声を機械に認識させる音声認識装置において、特徴量抽出部が、分析対象音声から、所定の周期毎に設定された一定長の複数の時間窓から音声特徴量を抽出する技術が開示されている。この技術では、音声の周波数に関する周波数軸系特徴パラメータと、音声の振幅に関するパワー系特徴パラメータとを、それぞれ異なる周期で抽出する。
【０００４】
また、特開平５－３０３３９１号公報では、特徴パラメータを計算するための単位時間（フレーム）を複数個用意する、あるいは各音韻毎に用意し、各フレーム長毎に特徴パラメータ時系列を計算し、そのそれぞれに対して音韻照合を行い、最適なものを選ぶ技術が開示されている。
【特許文献１】特開２００１－３５６７９０号公報明細書
【特許文献２】特開平５－３０３３９１号公報明細書
【発明の開示】
【発明が解決しようとする課題】
【０００５】
ただし、上述したような、一定長の複数の時間窓を一定時間ごとにずらしながら音声を特徴量の時系列に変換するという方法では、音韻の長さに応じて、抽出される特徴パラメータの数が異なってしまう。その結果、この特徴パラメータの数が、認識率に影響を与えてしまうという問題があった。
【０００６】
本発明は、上記のような問題を解決するためになされたものであって、その目的は、各音韻の認識率を向上することが可能な特徴パラメータの計算方法を用いた音声認識装置を提供することである。
【課題を解決するための手段】
【０００７】
このような目的を達成するために、本発明の音声認識装置は、入力音声信号に対して、各々が所定時間長を有する時間窓に対応する複数のフレームを、少なくとも異なる時間幅でスライドさせることにより、特徴パラメータの抽出を行う特徴抽出手段と、入力音声の音韻パターンにそれぞれ対応づけて標準パターンデータを格納するための記憶手段と、特徴抽出手段で抽出された特徴パラメータと標準パターンデータを照合することで、対応する音韻を認識し、認識結果を出力するための認識手段とを備える。
【発明の効果】
【０００８】
本発明の音声認識装置では、音韻の平均継続時間長が長い場合でも、あるいは、短い場合でも、処理負荷を低減しつつ、各音韻の認識率を向上することが可能である。
【発明を実施するための最良の形態】
【０００９】
以下、図面を参照して本発明の実施の形態について説明する。
【００１０】

（本発明の構成の説明の前提）

以下では、まず、本発明の音声認識装置の構成を説明する前提として、一般的な音声認識装置１０の構成および動作について、予め説明しておく。

【００１１】

図１は、このような音声認識装置１０の構成を説明するための機能ブロック図である。

【００１２】

図１を参照して、特徴検出部１０２は、入力として与えられた入力音声１０１に対して、ＬＰＣケプストラム係数（数十ミリ秒の音声切り出し単位であるフレームごとの対数パワースペクトル包絡のフーリエ変換）等の特徴パラメータを計算する。すなわち、特徴検出部１０２は、特徴量を計算する場合、通常数ミリ秒から数十ミリ秒を単位時間（フレーム）とし、１フレームの時間内では特徴量すなわち音声の波の構造は定常状態にあると近似して、特徴パラメータを計算する。そして、フレームをある時間だけずらして（これをフレームシフトという）、ふたたび特徴パラメータを計算する。これを繰り返すことによって特徴パラメータの時系列が得られる。

【００１３】

認識部１０３は、このようにして得られた特徴パラメータの時系列を、記憶装置に格納された単語辞書データベース（単語辞書ＤＢ）１０４内の標準パターンと比較し、類似度を計算することによって、認識結果１０５を出力する。

【００１４】

図２は、図１に示した特徴検出部１０２におけるフレームシフトを説明するための概念図である。

【００１５】

図２に示すように、音声認識装置１０の特徴検出部１０２では、フレームシフトの時間幅Ｄ２０１は、一定である。このため、音韻の継続時間の長い単語と、短い単語で、特徴パラメータの数が異なることになる。したがって、音韻の長い単語は、認識率が良く、音韻の短い単語は、認識率が音韻の長い単語に比べて悪いという傾向が生じうる。

【００１６】

本発明においては、以下に説明するように、フレームシフトの時間幅を可変にして、特徴パラメータを計算することで、音韻の解析を左右するとされる箇所に重点をおいて、音韻の長い単語と、音韻の短い単語とで、特徴パラメータの生成数を同数とするように特徴量の抽出を行う。

【００１７】

［実施の形態１］

以下、本発明の実施の形態１の音声認識装置１００の構成および動作について説明する。

【００１８】

図３は、音声認識装置１００の構成を説明するための機能ブロック図である。

【００１９】

図３に示した音声認識装置１００の構成は、基本的には、図１に示した音声認識装置１０の構成と同様である。

【００２０】

ただし、以下に説明するとおり、まず、発話者の音声がデジタル化されたものである入力音声３０１を受ける特徴抽出部３０２において、特徴パラメータ計算部３０２１は、フレームシフトの間隔を音韻の語頭のフレーム間隔ほど密にし、語尾に向けて順次フレーム間隔を粗くすることで、特徴パラメータを計算する。さらに、このようにして計算された特徴パラメータの時系列を受けて認識処理部３０３が認識処理を行うにあたり、参照する単語辞書データベース３０４が、このような可変なフレーム間隔に対応するように、以下に説明するとおり、所定の規則で変化するフレーム間隔に応じた標準パターンを予め格納する構成となっている。認識処理部３０３は、このような単語辞書データベース３０４を参照して、特徴パラメータの時系列との照合を行って解析することにより、認識結果３０

５を出力する。

【００２１】

以下、音声認識装置１００の動作についてさらに詳しく説明する。

【００２２】

音韻認織をおこなう場合、それぞれの音韻の平均継続時間長が重要である。音韻の特徴は、大きく語頭、話中、語尾の３つに分けられる。発音記号の／t／や／r／で表される子音は、語頭・話中・語尾の平均継続時間長は１５ミリ秒程度しかないのに対し、母音の方はそれぞれ１００ミリ砂を越える平均時間長をもつ。このように継続時間長に大きなばらつきがある様々な音素を認識するにあたり、単語の先頭データの重要度が特に高い。このため、本発明では、フレームシフトの時間幅を、以下に説明する所定の規則に基づいて変化させる。

【００２３】

図４は、音声認識装置１００の特徴パラメータ計算部３０２１のフレームシフト動作を説明するための概念図である。

【００２４】

例えば、図４においては、サンプリング周波数２０キロヘルツで１６ビットで量子化された入力音声３０１から、特徴パラメータ計算部３０２１において、特徴パラメータを計算するものとする。

【００２５】

特徴パラメータ計算部３０２１は、時間窓である固定のフレーム長Ｌを、入力音声の先頭から、終わりまで、順次長くなっていく時間幅Ｄ３０１～Ｄ３０ｎ（例：Ｄ３０１＜Ｄ３０２＜Ｄ３０３＜…＜Ｄ３０ｎ、ｎ：自然数）でシフトし、
それぞれ特徴パラメータ時系列Ｓ１～Ｓｎを生成する。

【００２６】

ここで、特に限定されないが、たとえば、時間幅Ｄ３０１～Ｄ３０ｎを順次長くしていくにあたっては、たとえば、先頭のフレームから次のフレームまでの時間間隔Ｄ３０１を基準として、所定の割合で等比級数的に以後の時間間隔Ｄ３０２～Ｄ３０ｎを順次長くしていくことや、所定の間隔で等差級数的に以後の時間間隔Ｄ３０２～Ｄ３０ｎを順次長くしていくことが可能である。もちろん、より一般的に、時間に対して単調に増加する関数にしたがって、時間間隔Ｄ３０２～Ｄ３０ｎを順次長くしていくことも可能である。

【００２７】

まず、この入力音声３０１の先頭からフレーム長Ｌ分のデータに注目し、この中のデータを定常状態にあるとみなして特徴パラメータを計算する。たとえば、１２次の線形予測係数ＬＰＣ（Linear Predictive Coding）から１６次のＬＰＣケプストラム係数を計算して１６次元の特徴ベクトルとする。次にフレームを時間幅Ｄ３０ｉ（ｉ＝１～ｎ）でシフトさせ、同様に特徴ベクトルを計算する。この換作を音声入力３０１の終わりまでくりかえすことによって、固定のフレーム長Ｌを用いて計算した特徴パラメータ時系列Ｓｎが得られる。

【００２８】

特徴パラメータ計算部３０２１からの特徴パラメータの出力に対しては、認識処理部３０３において、フレームごとに、単語辞書データベース３０４とのパラメータ比較が行われる。全フレームの比較が行われ、単語辞書データベース３０４に登録されたモデルのうちで、しきい値を満たす最も適するものが、認織結果３０５として出力される。

【００２９】

このとき、単語辞書データベース３０４へ格納するデータとしては、予め個々の音韻モデルに関して、フレーム長Ｌにおいて時間幅Ｄ３０１～Ｄ３０ｎでフレームシフトして計算した特徴パラメータを用いて、標準パターンを作成しておく。このような標準パターンは、あらかじめ発話内容と音韻の区間が既知の音声データベースを用い、計算した特徴パラメータ時系列を、個別の隠れマルコフモデル（ＨＭＭ：Hidden Markov Model）Ｐ０１を用意してトレーニングすることによって作成される。こうして、得られた音韻数Ｍ（Ｍ

：所定の自然数）の隠れマルコフモデルにより、単語辞書データベース１０４が構成される。

【００３０】

認識処理部３０４では、音韻認識を行うにあたり、すべての音韻の存在位置・存在確率を調べ、存在位置が重なっているものに関しては存在確率の大きなもののみを残す。こうして得られた音韻列を認識結果１０５として出力するものとする。

【００３１】

以上のような構成を有する音声認識装置１００により、フレームシフトの時間幅を固定した場合の音韻認識率と比較して、音韻の語頭に対する特徴パラメータの重み付けを大きくすることで、認識率を向上させることが可能となる。

【００３２】

［実施の形態２］

図５は、実施の形態２の音声認識装置２００の構成を説明するための機能ブロック図である。

【００３３】

なお、以下では、図２に示したように、時間窓であるフレーム間の間隔を固定して、特徴パラメータを抽出する処理手順を「固定フレーム間隔抽出処理」と呼ぶこととする。

【００３４】

図５に示した音声認識装置２００は、デジタル化された入力音声４０１に対して、第１の時間間隔での固定フレーム間隔抽出処理を行う第１の特徴パラメータ計算部を有する第１の特徴抽出部４０２と、第２の時間間隔での固定フレーム間隔抽出処理を行う第２の特徴パラメータ計算部を有する第２の特徴抽出部４０３とを備える。

【００３５】

第１の特徴抽出部４０２および第２の特徴抽出部４０３にて、それぞれ第１の特徴パラメータ時系列Ｓ０１～Ｓ０ｎおよび第２の特徴パラメータ時系列Ｓ１１～Ｓ１ｎを計算する。

【００３６】

また、音声認識装置２００は、予め第１の時間間隔での固定フレーム間隔抽出処理に対応した音韻モデルが登録された第１の単語辞書データベース４０２２と、予め第２の時間間隔での固定フレーム間隔抽出処理に対応した第２の単語辞書データベース４０３２と、第１の特徴抽出部４０２で計算された特徴パラメータのそれぞれを第１の単語辞書データベース４０２２内のデータと比較して音韻を認識するための第１の認識処理部４０２１と、第２の特徴抽出部４０３で計算された特徴パラメータのそれぞれを第２の単語辞書データベース４０３２内のデータと比較して音韻を認識するための第２の認識処理部４０３１と、さらに、第１および第２の認識処理部４０２１，４０３１の認識結果を、その適合度に応じて選択し、認識結果４０５を得るための結果選択部４０４とを備える。

【００３７】

以下、音声認識装置２００の動作について、さらに詳しく説明する。

【００３８】

まず、入力音声４０１の先頭からフレーム長Ｌ分のデータに注目し、この中のデータを定常状態にあるとみなして、第１の特徴抽出部４０２および第２の特徴抽出部４０３にて、特徴パラメータを計算する。

【００３９】

音声認識装置２００では、第１の特徴抽出部４０２において、たとえば、１２次の線形予測係数ＬＰＣから１６次のＬＰＣケプストラム係数を計算して１６次元の特徴ベクトルとする。同様に、第２の特徴抽出部４０３においても、１２次の線形予測係数ＬＰＣから１６次のＬＰＣケプストラム係数を計算して１６次元の特徴ベクトルとする。

【００４０】

その結果、第１の特徴抽出部４０２および第２の特徴抽出部４０３のそれぞれにおいて、第１の特徴パラメータＳ０１、第２の特徴パラメータＳ１１が得られる。この操作以降

、入力音声４０１の信号の終わりまで、第１の特徴抽出部４０２では、固定時間幅Ｄ２０１でフレームシフトを繰り返し計算した第１の特徴パラメータＳ０ｎを出力し、第２の特徴抽出部４０３では、固定時間幅Ｄ２０１１（＜Ｄ２０１）でフレームシフトを繰り返し計算した第２の特徴パラメータＳ１ｎを出力する。

【００４１】

一方、あらかじめ個々の音韻モデルに関して、フレーム長Ｌから計算した特徴パラメータを用いて、第１の標準パターンを作成しておく。この第１の標準パターンは、あらかじめ発話内容と音韻の区間が既知の音声データベースを用いて計算した特徴パラメータ時系列（ここで、この特徴パラメータ時系列は、フレームシフトの時間幅をＤ２０１にして、作成したものである）を、個別の隠れマルコフモデル（ＨＭＭ）Ｐ０１を用意してトレーニングすることによって作成しておくものとする。こうして得られた音韻数Ｍの隠れマルコフモデルにより、第１の単語辞書データベース４０２２が構成される。

【００４２】

また、第２の標準パターンも同様に、あらかじめフレーム長Ｌから計算した特徴パラメータを用いて、作成しておく。この第２の標準パターンは、あらかじめ発話内容と音韻の区間が既知の音声データベースを用い、計算した特徴パラメータ時系列（ここで、この特徴パラメータ時系列は、フレームシフトの時間幅をＤ２０１１にして、作成したものである）を、個別の隠れマルコフモデル（ＨＭＭ）Ｐ１１を用意してトレーニングすることによって作成しておくものとする。こうして得られた音韻数Ｍの隠れマルコフモデルにより、第２の単語辞書データベース４０３２が構成される。

【００４３】

第１の認識処理部４０２１においては、入力音声の先頭のフレームから順に各音韻毎に特徴パラメータ時系列Ｓ０１は標準パターンＰ０１を用い、特徴パラメータ時系列Ｓ０２には標準パターンＰ０２を用いて照合を行ない、以下同様にして、特徴パラメータ時系列Ｓ０ｎには標準パターンＰ０ｎを用いて音韻照合を行い、存在位置および存在確率の重なるものを出力する。

【００４４】

同様に、第２の認識処理部４０３１においては、入力音声の先頭のフレームから順に各音韻毎に特徴パラメータ時系列Ｓ１１は標準パターンＰ１１を用い、特徴パラメータ時系列Ｓ１２には標準パターンＰ１２を用いて照合を行ない、以下同様にして、特徴パラメータ時系列Ｓ１ｎには標準パターンＰ１ｎを用いて音韻照合を行い、存在位置および存在確率の重なるものを出力する。

【００４５】

結果選択部４０４では、第１の認識処理部４０２１および第２の認識処理部４０３１から出力されたすべての音韻の存在位置・存在権率を調べ、存在位置が重なっているものに関しては存在確率の大きなもののみを残す。結果選択部４０４は、こうして得られた音韻列を認識結果４０５として出力する。

【００４６】

以上説明したような音声認識装置２００の構成により、フレーム間の時間間隔を固定した場合の音韻認識率と比較して、異なったフレーム間の時間間隔で抽出された特徴パラメータを用いて、より存在確率の高い方が選択されるので、認識率を向上することができる。

【００４７】

［実施の形態３］

以下では、図４で説明したように、時間窓であるフレーム間の間隔を順次長くしながら、特徴パラメータを抽出する処理手順を「可変フレーム間隔抽出処理」と呼ぶこととする。

【００４８】

実施の形態２では、第１の特徴抽出部４０２と第２の特徴抽出部４０３との双方が、固定フレーム間隔抽出処理を行なうものとした。

【００４９】
これに対して、本発明の実施の形態３の音声認識装置の基本的な構成は、実施の形態２の音声認識装２００の構成と同様である。
【００５０】
ただし、実施の形態３の音声認識装置では、第２の特徴抽出部４０３は、可変フレーム間隔抽出処理を行なうものとする。
【００５１】
すなわち、第２の特徴抽出部４０３は、図４で説明したようにフレームシフトの時間幅Ｄ３０ｉ（ｉ：自然数、Ｄ３０１＜Ｄ３０２＜Ｄ３０３＜…）を順次長くしながら可変にし、特徴パラメータをそれぞれにおいて計算する。
【００５２】
また、第２の単語辞書データベース４０３２には、フレームシフトの時間幅をＤ３０ｉ（ｉ：自然数、Ｄ３０１＜Ｄ３０２＜Ｄ３０３＜…）にして計算した特徴パラメータを用いて、標準パターンを作成しておくものとする。
【００５３】
実施の形態３の音声認識装置のその他の構成は、実施の形態２の音声認識装置２００の構成と同様であるので、その説明は繰り返さない。
【００５４】
このような実施の形態３の音声認識装置の構成により、音声認識装置２００の奏する効果に加え、音韻の平均継続時間長が長い場合は、固定フレーム間隔抽出処理で有効に対処することが可能であり、一方、音韻の平均継続時間長が短い場合は、可変フレーム間隔抽出処理で有効に対処することが可能であるので、処理負荷を低減できる。
【００５５】
[実施の形態４]
図６は、実施の形態４の音声認識装置３００の構成を説明するための機能ブロック図である。
【００５６】
図６に示した音声認識装置３００は、デジタル化された入力音声５０１に対して、第１の時間間隔での固定フレーム間隔抽出処理を行う第１の特徴パラメータ計算部を有する第１の特徴抽出部５０２と、第２の時間間隔での固定フレーム間隔抽出処理を行う第２の特徴パラメータ計算部を有する第２の特徴抽出部５０３とを備える。
【００５７】
さらに、音声認識装置３００は、後に説明する制御信号５１を入力とするインバータ５１１と、制御信号５１およびインバータ５１１の出力信号５０に応じて、入力音声５０１を、第１の特徴抽出部５０２または第２の特徴抽出部５０３に選択的に与えるための入力選択部５１０を備える。
【００５８】
入力選択部５１０は、入力音声５０１および制御信号５１を入力に受け、出力を第１の特徴抽出部５０２に与えるＡＮＤ回路５１２と、入力音声５０１およびインバータ５１１の出力信号５０を入力に受け、出力を第２の特徴抽出部５０３に与えるＡＮＤ回路５１３とを備える。
【００５９】
第１の特徴抽出部５０２および第２の特徴抽出部５０３にて、それぞれ第１の特徴パラメータ時系列Ｓ０１～Ｓ０ｎおよび第２の特徴パラメータ時系列Ｓ１１～Ｓ１ｎを計算する。
【００６０】
また、音声認識装置３００は、予め第１の時間間隔での固定フレーム間隔抽出処理に対応した音韻モデルが登録された第１の単語辞書データベース５０２２と、予め第２の時間間隔での固定フレーム間隔抽出処理に対応した第２の単語辞書データベース５０３２と、第１の特徴抽出部５０２で計算された特徴パラメータのそれぞれを第１の単語辞書データ

ベース５０２２内のデータと比較して音韻を認識するための第１の認識処理部５０２１と、第２の特徴抽出部５０３で計算された特徴パラメータのそれぞれを第２の単語辞書データベース５０３２内のデータと比較して音韻を認識するための第２の認識処理部５０３１と、さらに、第１および第２の認識処理部５０２１，５０３１の認識結果を、以下に説明する手順にしたがって選択し、認識結果５０５を得るための結果選択部５０４とを備える。

【００６１】

結果選択部５０４は、第１の認識処理部５０２１の出力および制御信号５１を入力に受け、認識結果５０５を出力するＡＮＤ回路５１４と、第２の認識処理部５０３１の出力および出力信号５０を入力に受け、認識結果５０５を出力するＡＮＤ回路５１５とを備える。

【００６２】

以下、音声認識装置３００の動作について説明する。

【００６３】

まず、入力音声５０１の先頭からフレーム長Ｌ分のデータに注目し、この中のデータを定常状態にあるとみなし、制御信号５１に応じて第１の特徴抽出部５０２、あるいは第２の特徴抽出部５０３において特徴パラメータを計算する。

【００６４】

ここで、制御信号５１は、第１の認識処理部５０２１における認識処理で、認識結果を得るために設定したしきい値を満たす場合は、第１の特徴抽出部５０２に音声を入力し、第１の認識処理部５０２１ではしきい値を満たさない場合には、第２の特徴抽出部５０３に音声を入力するように変化するものとする。

【００６５】

例えば、入力音声５０１が、登録単語のいくつかと、語頭は同じであるが、語尾になると異なるような場合、第１の特徴抽出部５０２および第１の認識処理部５０２１からなる第１の処理系で、語頭から語尾にかけてフレームごとに認識処理を行うにつれて、次第にしきい値を満たさなくなっていくことが起り得る。

【００６６】

このとき、第１の認識処理部５０２１は、制御フラグを制御信号５１として返し、そのフラグによって、第２の特徴抽出部５０３および第２の認識処理部５０３１からなる第２の処理系に認識処理を切り替え、シフト時間幅を変化させて認識処理を行うものとする。

【００６７】

実施の形態４では、上述した第２の処理系でフレームシフトの時間幅が、第１の処理系でのフレームシフトの時間幅よりも短いものであることとして、以下説明する。

【００６８】

実施の形態４において、第１の特徴抽出部５０２および第２の特徴抽出部５０３においては、１２次の線形予測係数ＬＰＣから１６次のＬＰＣケプストラム係数を計算して１６次元の特徴ベクトルとするものとする。

【００６９】

その結果、第１の特徴抽出部５０２および第２の特徴抽出部５０３のそれぞれにおいて、第１の特徴パラメータＳ０１、第２の特徴パラメータＳ１１が得られる。この操作以降、入力信号の終わりまで、第１の特徴抽出部５０２では、一定値に固定した時間幅Ｄ２０１でフレームシフトを繰り返し、計算した第１の特徴パラメータＳ０ｎを出力し、第２の特徴抽出部５０３では、固定時間幅Ｄ２０１１（＜Ｄ２０１）でフレームシフトを繰り返し計算した第２の特徴パラメータＳ１ｎを出力する。

【００７０】

また、第１および第２の単語辞書データベース５０２２および５０３２には、実施の形態２と同様にして、フレームシフトの時間幅をＤ２０１にして作成した特徴パラメータ時系列およびフレームシフトの時間幅をＤ２０１１にして作成した特徴パラメータ時系列のそれぞれに対応した、各音韻モデルに対する隠れマルコフモデルによる第１および第２の

標準パターンが格納されているものとする。

【００７１】

第１の認識処理部５０２１においては、入力音声の先頭のフレームから順にフレーム毎に、特徴パラメータ時系列Ｓ０１には標準パターンＰ０１を用い、特徴パラメータ時系列Ｓ０２には標準パターンＰ０２を用いる。以下同様に、第１の認識処理部５０２１は、特徴パラメータ時系列Ｓ０ｘには標準パターンＰ０ｘ（ｘ：自然数）を用い、存在位置、存在確率の重なり、設定するしきい値を満たすものを出力する。この処理を繰り返す中で、設定したしきい値を満たさなければ、第１の認識処理部５０２１は、切り替え信号を生成して制御倍号５１を反転させ、第２の特徴抽出部５０３の出力を用いて、第２の認識処理部５０３１において音韻照合を行なうように処理を切替える。すなわち、以後、第２の認識処理部５０３１は、同様に、フレーム毎に特徴パラメータ時系列Ｓ１（ｘ＋１）には標準パターンＰ１（ｘ＋１）を用い、特徴パラメータ時系列Ｓ１（ｘ＋２）には標準パターンＰ１（ｘ＋２）を用い、以下同様にして、特徴パラメータ時系列Ｓ１ｎには標準パターンＰ１ｎを用いて音韻照合を行い、存在位置、存在確率の重なるものを出力する。

【００７２】

そして、結果選択部５０４は、第１または第２の処理系の結果から得られた音韻列を最終的な認識結果５０５として出力する。

【００７３】

以上説明したような実施の形態４の音声認識装置３００の構成により、フレームの時間幅を単一に固定した場合の音韻認識率と比較して、認識率を向上させることが可能である。

【００７４】

なお、もう一つの効果として、例えば、図示しないもう一つ別の処理系があり、その処理系は特定のものとは限定しないが、その図示しない処理系が処理中であるということを示す信号を生成できるものとし、その生成信号を制御信号５１として使用することも可能である。その場合、本音声信号処理装置３００を含むシステムにおいて、ＣＰＵ（Central Processing Unit）などの処理負荷を低減できる。

【００７５】

［実施の形態５］

実施の形態４では、第１の特徴抽出部５０２と第２の特徴抽出部５０３との双方が、固定フレーム間隔抽出処理を行なうものとした。

【００７６】

これに対して、本発明の実施の形態５の音声認識装置の基本的な構成は、実施の形態４の音声認識装３００の構成と同様である。

【００７７】

ただし、実施の形態５の音声認識装置では、第２の特徴抽出部５０３は、可変フレーム間隔抽出処理を行なうものとする。

【００７８】

すなわち、第２の特徴抽出部５０３は、図４で説明したようにフレームシフトの時間幅Ｄ３０ｉ（ｉ：自然数、Ｄ３０１＜Ｄ３０２＜Ｄ３０３＜…）を順次長くしながら可変にし、特徴パラメータをそれぞれにおいて計算する。

【００７９】

また、第２の単語辞書データベース５０３２には、フレームシフトの時間幅をＤ３０ｉ（ｉ：自然数、Ｄ３０１＜Ｄ３０２＜Ｄ３０３＜…）にして計算した特徴パラメータを用いて、標準パターンを作成しておくものとする。

【００８０】

実施の形態５の音声認識装置のその他の構成は、実施の形態４の音声認識装３００の構成と同様であるので、その説明は繰り返さない。

【００８１】

このような実施の形態５の音声認識装置の構成により、音声認識装置３００の奏する効

果に加え、音韻の平均継続時間長が長い場合は、固定フレーム間隔抽出処理で有効に対処することが可能であり、一方、音韻の平均継続時間長が短い場合は、可変フレーム間隔抽出処理で有効に対処することが可能であるので、処理負荷を低減できる。

【００８２】

［実施の形態６］

図７は、実施の形態６の音声認識装置４００の構成を説明するための機能ブロック図である。

【００８３】

図７に示した音声認識装置４００においては、入力音声６０１、入力選択部６１０、制御信号６１、インバータ６１１、第１の特徴抽出部６０２、第２の特徴抽出部６０３、第１の認識処理部６０２１、第２の認識処理部６０３１、結果選択部６０４、第１の単語辞書データベース６０２２および認識結果６０５は、それぞれ、実施の形態４の音声認識装置３００の入力音声５０１、入力選択部５１０、制御信号５１、インバータ５１１、第１の特徴抽出部５０２、第２の特徴抽出部５０３、第１の認識処理部５０２１、第２の認識処理部５０３１、結果選択部５０４、第１の単語辞書データベース５０２２および認識結果５０５に相当する機能を有している。

【００８４】

図７に示した音声認識装置４００においては、実施の形態４の音声認識装置３００の構成とは異なり、第２の単語辞書データ５０３２の代わりに、データ補間部６０３２が設けられている。

【００８５】

図７に示した音声認識装置４００においても、第２の特徴抽出部５０３および第２の認識処理部５０３１からなる第２の処理系でのフレームシフトの時間幅Ｄ２０１１が、第１の特徴抽出部５０２および第１の認識処理部５０２１からなる第１の処理系でのフレームシフトの時間幅Ｄ２０１よりも短いものであるとする。

【００８６】

ここで、音声認識装置４００においても、あらかじめ個々の音韻モデルに関して、フレーム長Ｌから計算した特徴パラメータを用いて、第１の標準パターンを作成しておく。この第１の標準パターンは、あらかじめ発話内容と音韻の区間が既知の音声データベースを用いて計算した特徴パラメータ時系列（ここで、この特徴パラメータ時系列は、フレームシフトの時間幅をＤ２０１にして、作成したものである）を、個別の隠れマルコフモデル（ＨＭＭ）Ｐ０１を用意してトレーニングすることによって作成しておくものとする。こうして得られた音韻数Ｍの隠れマルコフモデルにより、第１の単語辞書データベース６０２２が構成される。

【００８７】

図８は、このようにして作成された標準パターンが、第１の単語辞書データベース６０２２に格納される状態を説明するための概念図である。

【００８８】

図８に示すとおり、音韻に対応した隠れマルコフモデルに対して、所定の時間における８０１～８０ｎの第１の標準パターンは、それぞれ時刻ｔ１～ｔｎにおけるパラメータｍ１～ｍｎとして構成される。

【００８９】

音声認識装置４００では、第２の処理系でのフレームシフトの時間幅Ｄ２０１１が、第１の処理系でのフレームシフトの時間幅Ｄ２０１よりも短いものであることから、第２の認識処理部５０３１で使用されるべき第２の標準パターンとして第１の標準パターンを用いようとしたとしても、第１の単語辞書データベース６０２２には、第２の標準パターンとしては存在しない部分が生じる。

【００９０】

そこで、音声認識装置４００では、第２の標準パターンを第１の標準パターンに基づいて、データ補間部６０３２により生成する。

【００９１】
図９は、データ補間部６０３２の処理を説明するための概念図である。
【００９２】
図９に示すように、第１の標準パターンと時間データを用いて中間データを線形補間（任意の高次関数でも可）によって計算することで、全ての時間における第２の標準パターンを作成できる。
【００９３】
音声認識装置４００のその他の動作は、実施の形態４と同様であるので、その説明は繰り返さない。
【００９４】
以上のような音声認識装置４００の構成とすれば、単語辞書データベースとして使用するメモリ等の記憶装置の記憶容量を削減できる。
【００９５】
［実施の形態７］
実施の形態６では、第１の特徴抽出部６０２と第２の特徴抽出部６０３との双方が、固定フレーム間隔抽出処理を行なうものとした。
【００９６】
これに対して、本発明の実施の形態７の音声認識装置の基本的な構成は、実施の形態６の音声認識装４００の構成と同様である。
【００９７】
ただし、実施の形態７の音声認識装置では、第２の特徴抽出部６０３は、可変フレーム間隔抽出処理を行なうものとする。
【００９８】
すなわち、第２の特徴抽出部６０３は、図４で説明したようにフレームシフトの時間幅Ｄ３０ｉ（ｉ：自然数、Ｄ３０１＜Ｄ３０２＜Ｄ３０３＜…）を順次長くしながら可変にし、特徴パラメータをそれぞれにおいて計算する。
【００９９】
また、第２の標準パターン生成においては、実施の形態６と同様に、第１の単語辞書データベース６０２２を用いて、データ補間部６０３２により、全ての標準パターンを生成する。
【０１００】
実施の形態７の音声認識装置のその他の構成は、実施の形態６の音声認識装４００の構成と同様であるので、その説明は繰り返さない。
【０１０１】
このような実施の形態７の音声認識装置の構成により、音声認識装置３００の奏する効果に加え、音韻の平均継続時間長が長い場合は、固定フレーム間隔抽出処理で有効に対処することが可能であり、一方、音韻の平均継続時間長が短い場合は、可変フレーム間隔抽出処理で有効に対処することが可能であるので、処理負荷を低減できる。
【０１０２】
［実施の形態８］
図１０は、実施の形態８の音声認識装置５００の構成を説明するための機能ブロック図である。
【０１０３】
図１０に示した音声認識装置５００の構成においては、入力音声７０１、入力選択部７１０、制御信号７１、インバータ７１１、第１の特徴抽出部７０２、第２の特徴抽出部７０３、第１の認識処理部７０２１、第２の認識処理部７０３１、結果選択部７０４、第１の単語辞書データベース７０２２および認識結果７０５は、それぞれ、実施の形態６の音声認識装置４００の入力音声６０１、入力選択部６１０、制御信号６１、インバータ６１１、第１の特徴抽出部６０２、第２の特徴抽出部６０３、第１の認識処理部６０２１、第２の認識処理部６０３１、結果選択部６０４、第１の単語辞書データベース６０２２およ

び認識結果６０５に相当する機能を有している。

【０１０４】

音声認識装置５００においても、第２の特徴抽出部７０３および第２の認識処理部７０３１からなる第２の処理系でのフレームシフトの時間幅Ｄ２０１１が、第１の特徴抽出部７０２および第１の認識処理部７０２１からなる第１の処理系でのフレームシフトの時間幅Ｄ２０１よりも長いものであるとする。

【０１０５】

音声認識装置５００では、時間幅の最小値はＤ２０１とする。

【０１０６】

音声認識装置５００においても、あらかじめ個々の音韻モデルに関して、フレーム長Ｌから計算した特徴パラメータを用いて、第１の標準パターンを作成しておく。この第１の標準パターンは、あらかじめ発話内容と音韻の区間が既知の音声データベースを用いて計算した特徴パラメータ時系列（ここで、この特徴パラメータ時系列は、フレームシフトの時間幅をＤ２０１にして、作成したものである）を、個別の隠れマルコフモデル（ＨＭＭ）Ｐ０１を用意してトレーニングすることによって作成しておくものとする。こうして得られた音韻数Ｍの隠れマルコフモデルにより、第１の単語辞書データベース７０２２が構成される。

【０１０７】

第１の第１の単語辞書データベース７０２２も、図８に示したように時刻とパラメータとが関連付けて格納されているものとする。

【０１０８】

音声認識装置５００では、第２の処理系でフレームシフトの時間幅Ｄ２０１１が、第１の処理系でのフレームシフトの時間幅Ｄ２０１よりも長いだけでなく、長い時間幅Ｄ２０１１で変化する際の各時刻点が、短い時間幅Ｄ２０１で変化する際の時刻点に相当または対応するように、時間幅Ｄ２０１１と時間幅Ｄ２０１との関係が定められているものとする。

【０１０９】

たとえば、時間幅Ｄ２０１１での変化に対して、時間幅Ｄ２０１の変化が、等比または等差的なものとする場合、第２の標準パターンは、実施の形態６のような特別な補間操作を必要とせずに、第１の標準パターンから得ることができる。

【０１１０】

実施の形態８の音声認識装置のその他の構成および動作は、実施の形態６の音声認識装置４００の構成と同様であるので、その説明は繰り返さない。

【０１１１】

このような実施の形態８の音声認識装置の構成により、音声認識装置４００の奏する効果に加え、一層、処理負荷を低減できる。

【０１１２】

今回開示された実施の形態はすべての点で例示であって制限的なものではないと考えられるべきである。本発明の範囲は上記した説明ではなくて特許請求の範囲によって示され、特許請求の範囲と均等の意味および範囲内でのすべての変更が含まれることが意図される。

【図面の簡単な説明】

【０１１３】

【図１】音声認識装置１０の構成を説明するための機能ブロック図である。

【図２】図１に示した特徴検出部１０２におけるフレームシフトを説明するための概念図である。

【図３】音声認識装置１００の構成を説明するための機能ブロック図である。

【図４】音声認識装置１００の特徴パラメータ計算部３０２１のフレームシフト動作を説明するための概念図である。

【図５】実施の形態２の音声認識装置２００の構成を説明するための機能ブロック図

である。

【図６】実施の形態４の音声認識装置３００の構成を説明するための機能ブロック図である。

【図７】実施の形態６の音声認識装置４００の構成を説明するための機能ブロック図である。

【図８】標準パターンが、第１の単語辞書データベース６０２２に格納される状態を説明するための概念図である。

【図９】データ補間部６０３２の処理を説明するための概念図である。

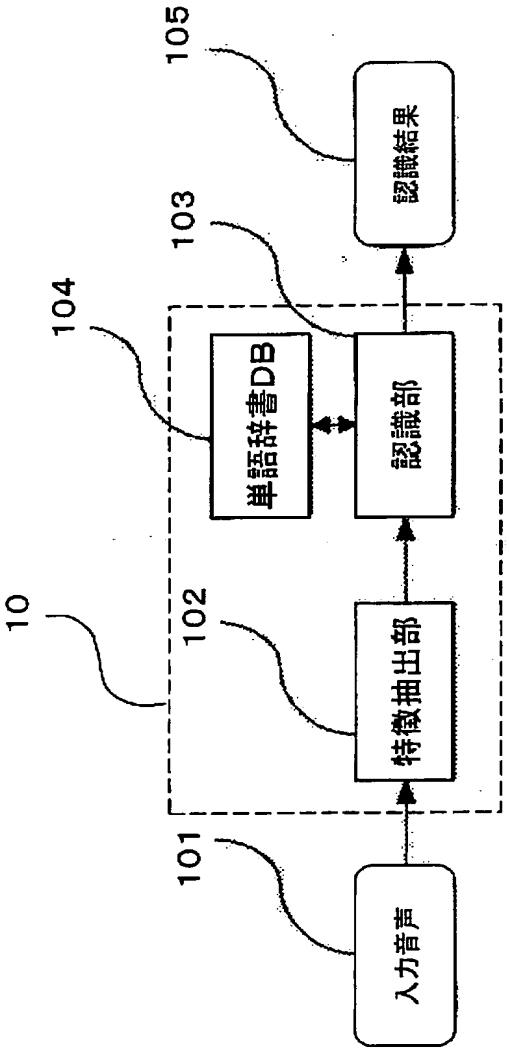
【図１０】実施の形態８の音声認識装置５００の構成を説明するための機能ブロック図である。

【符号の説明】

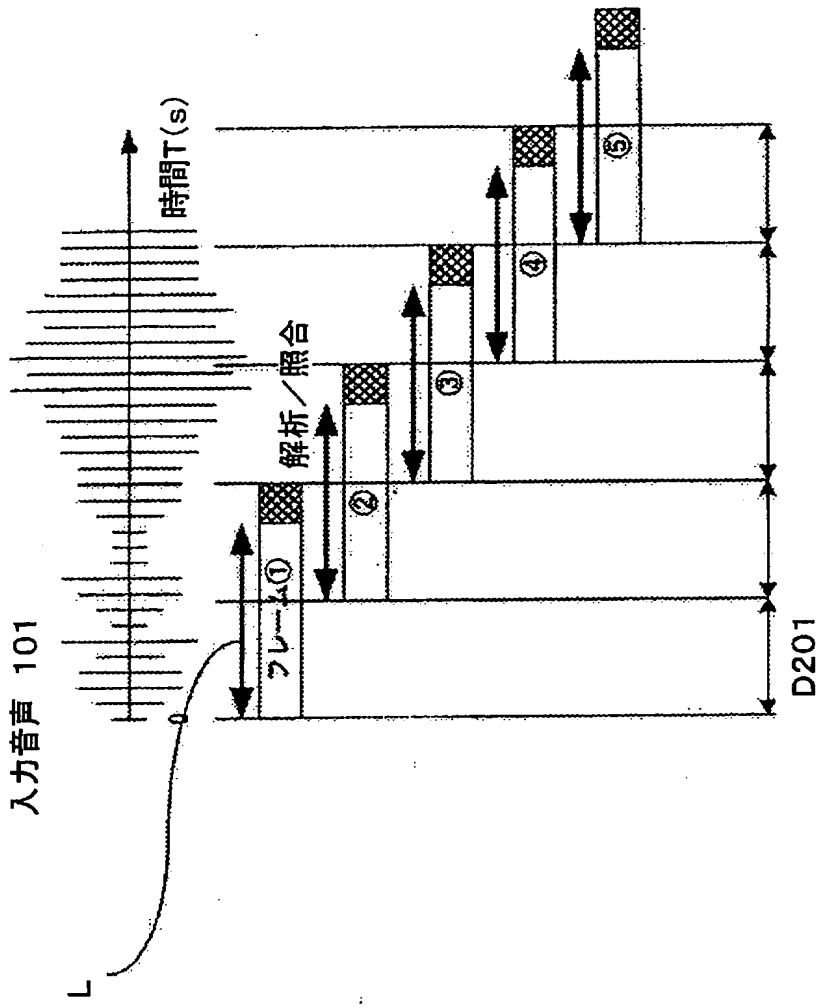
【０１１４】

５１，６１，７１　制御信号、１０１，３０１，４０１，５０１，６０１，７０１　入力音声、５１０，６１０，７１０　入力選択部、３０２　特徴抽出部、５１１，６１１，７１１　インバータ、４０２，５０２，６０２，７０２　第１の特徴抽出部、４０３，５０３，６０３，７０３　第２の特徴抽出部、４０２１，５０２１，６０２１，７０２１　第１の認識処理部、４０３１，５０３１，６０３１，７０３１　第２の認識処理部、４０４，５０４，６０４，７０４　結果選択部、４０２２，５０２２，６０２２，７０２２　第１の単語辞書データベース、４０３２，５０３２　第２の単語辞書データベース、６０３２　データ補間部、１０５，３０５，４０５，５０５，６０５，７０５　認識結果、１０、１００，２００，３００，４００，５００　音声認識装置。

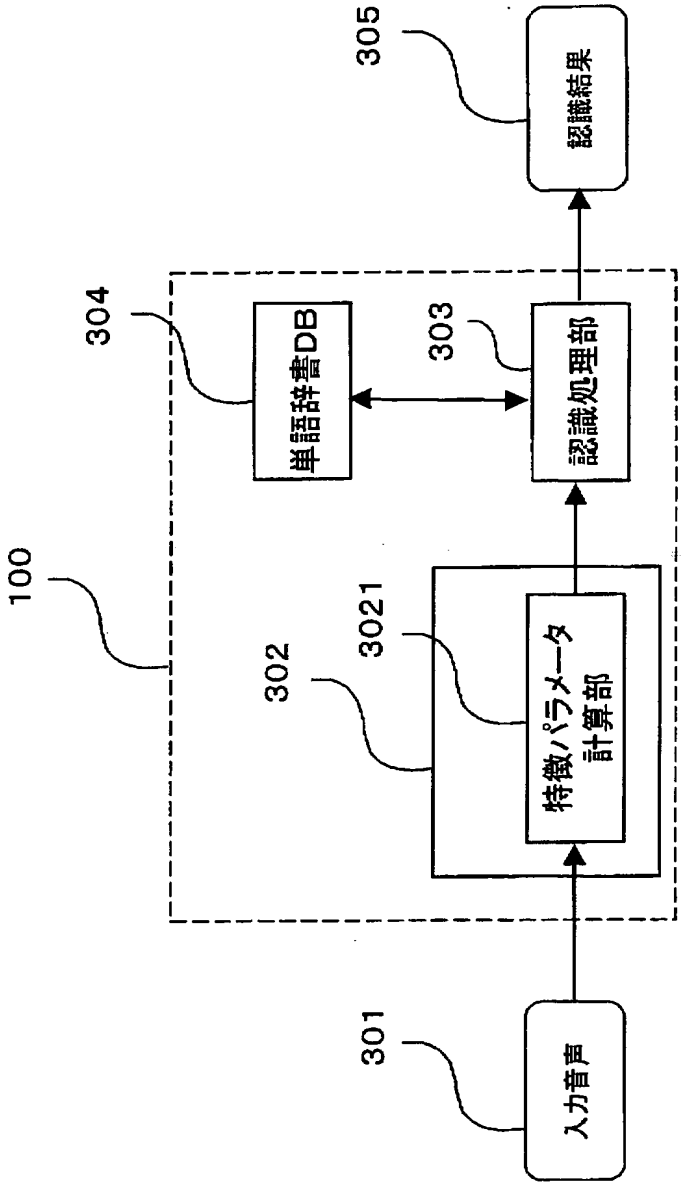
【書類名】図面
【図１】



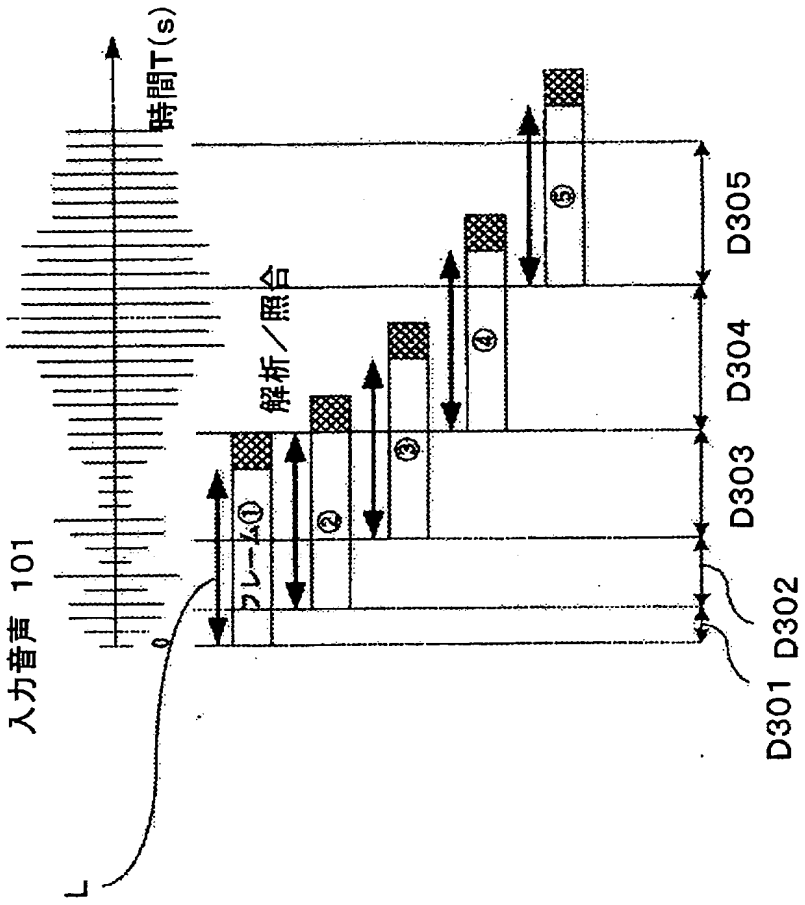
【図２】



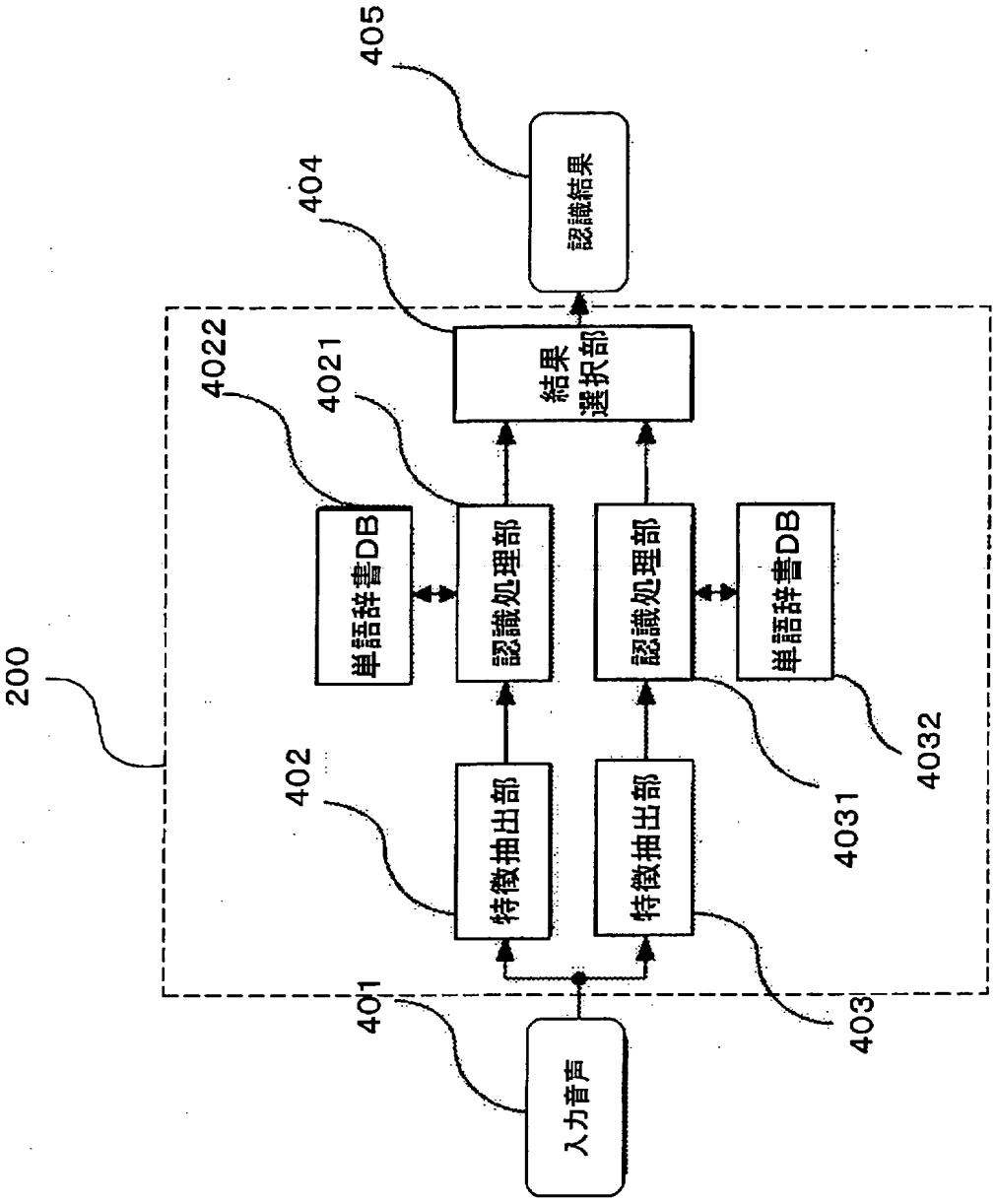
【図３】


305
認識結果
304
単語辞書DB
303
認識処理部
100
3021
302
特徴パラメータ
計算部
301
入力音声

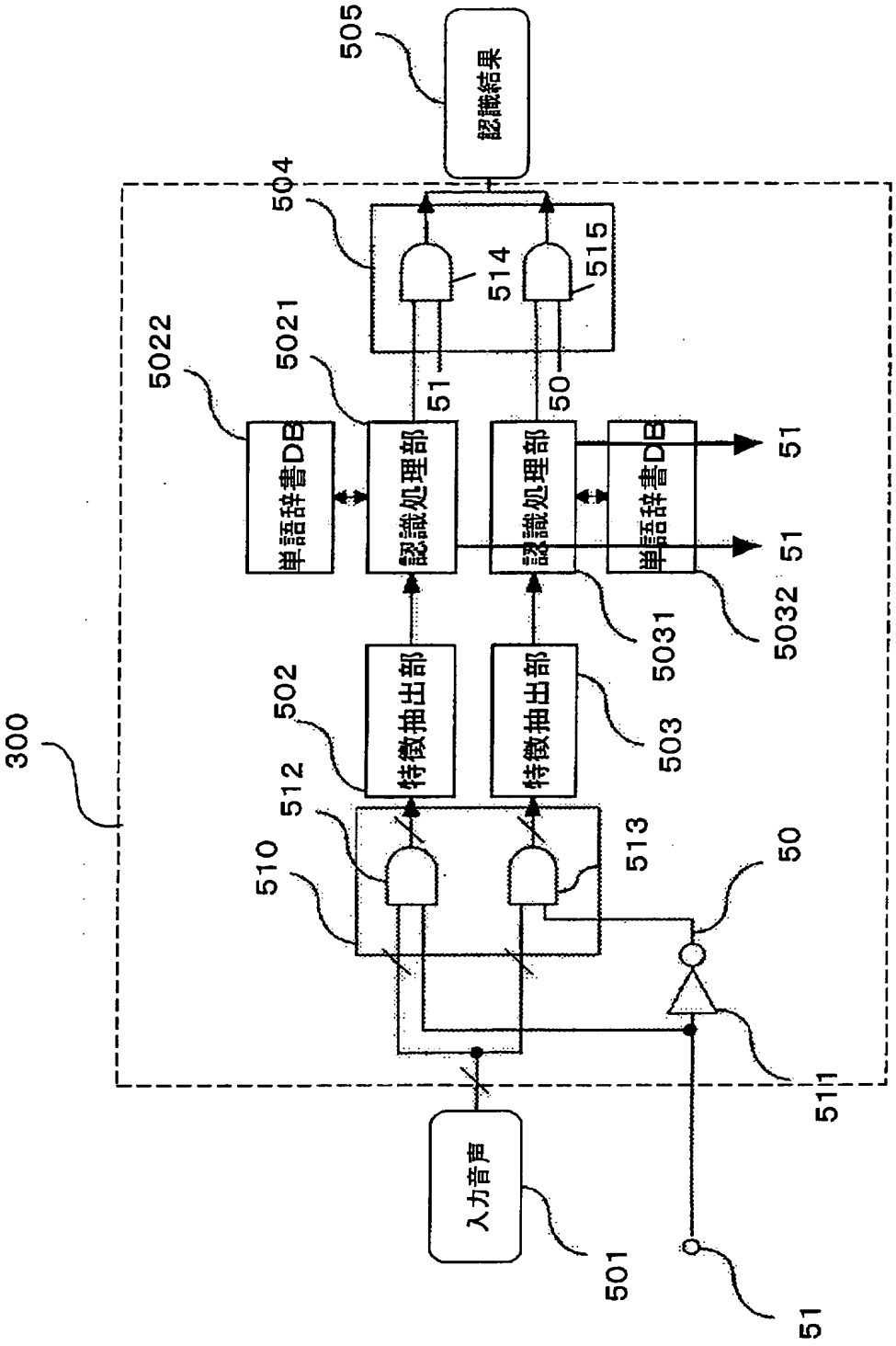
【図４】



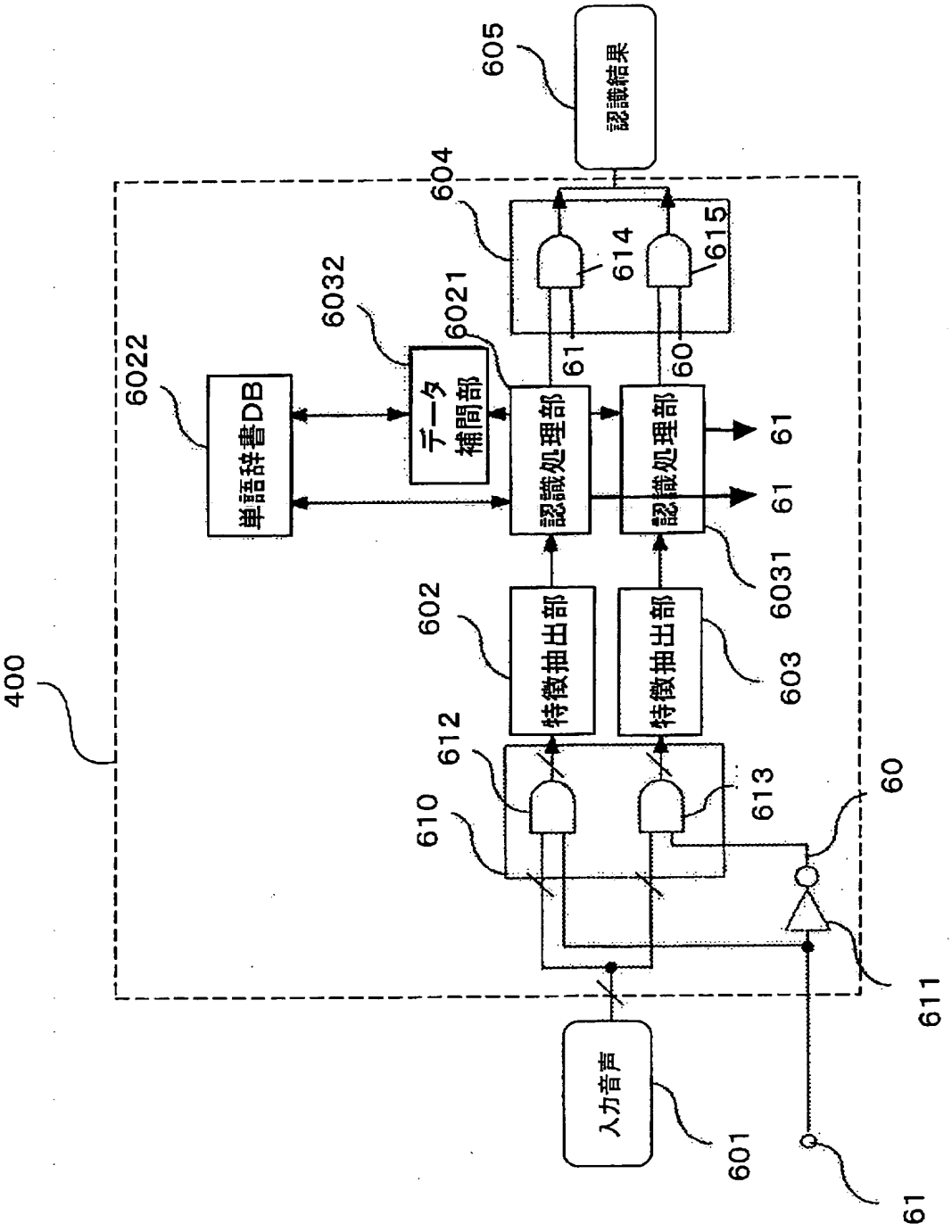
【図５】


405
認識結果
404
結果選択部
4022
4021
単語辞書DB
認識処理部
認識処理部
単語辞書DB
200
402
特徴抽出部
特徴抽出部
4031
4032
403
401
入力音声

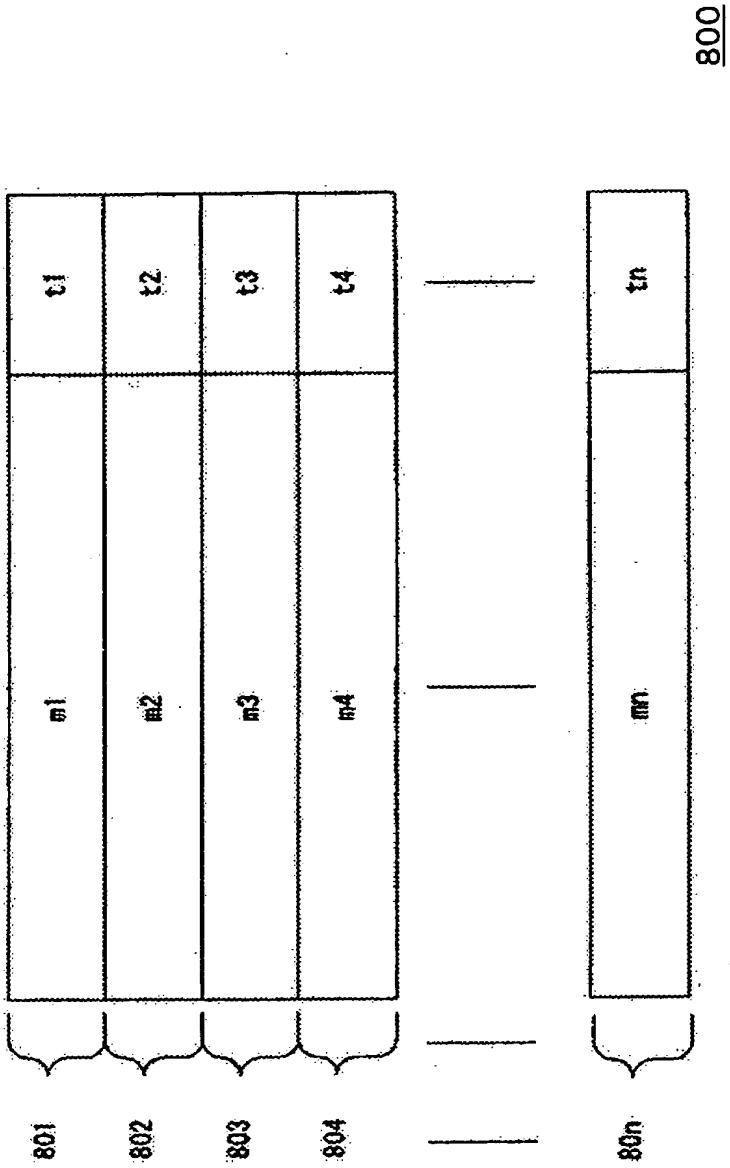
【図６】


505
認識結果
504
514
515
5022
5021
51
50
単語辞書DB
認識処理部
認識処理部
単語辞書DB
51
51
5032
5031
502
503
特徴抽出部
特徴抽出部
300
512
513
510
50
511
入力音声
501
51

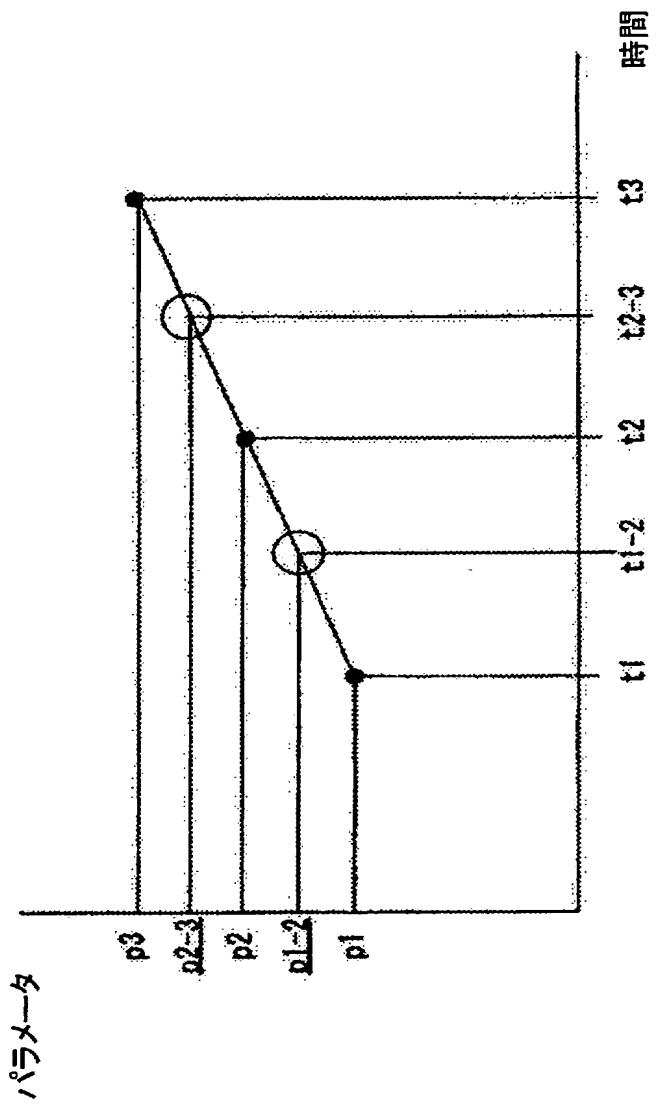
【図７】


400
605
認識結果
604
614
615
6032
6021
6022
単語辞書DB
データ補間部
61
60
認識処理部
認識処理部
61
61
6031
602
603
特徴抽出部
特徴抽出部
612
610
613
60
611
入力音声
601
61

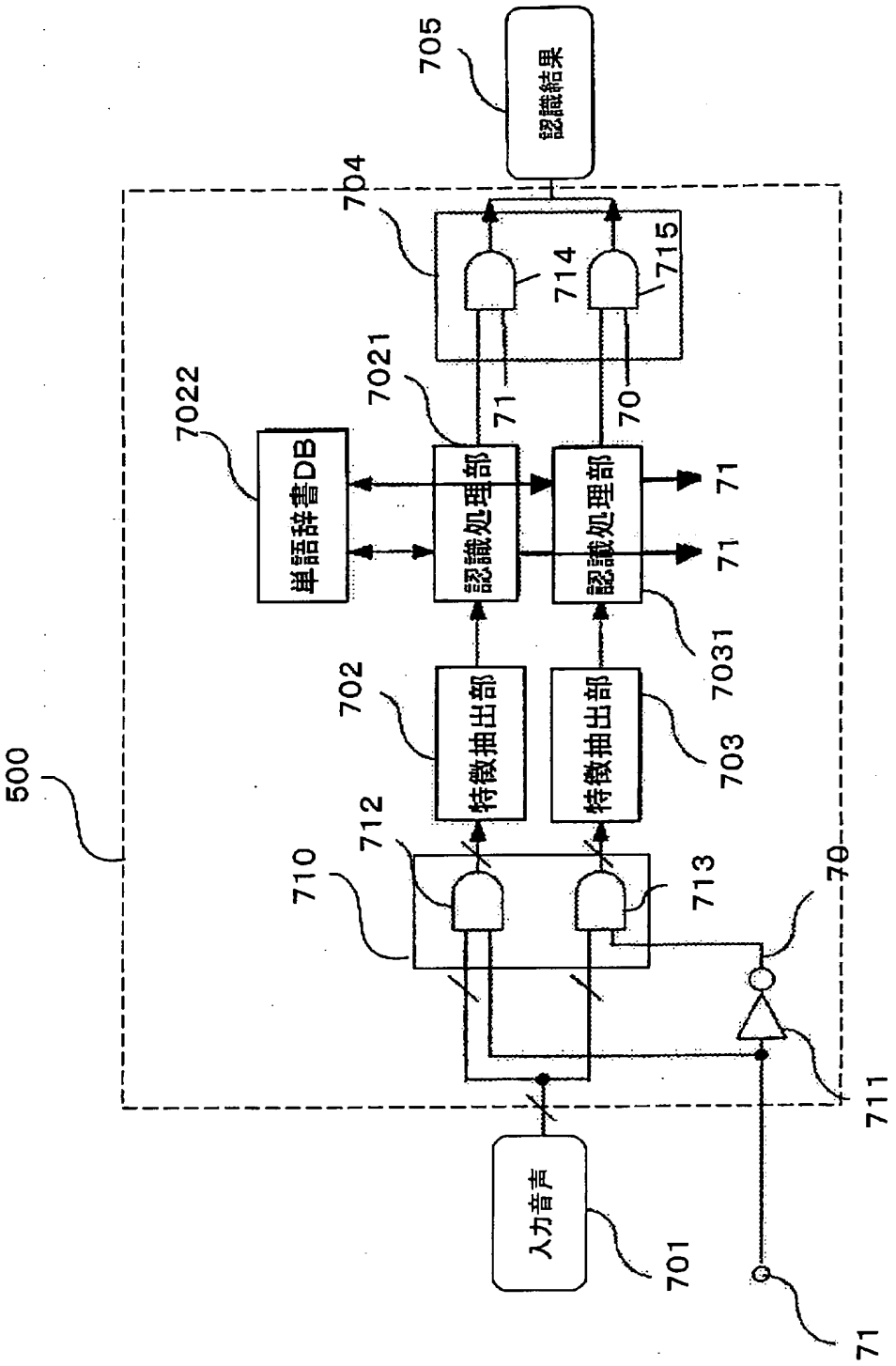
【図８】



【図９】



【図１０】


認識結果
705
704
714
715
7021
7022
単語辞書DB
認識処理部
認識処理部
71
70
71
71
7031
702
703
特徴抽出部
特徴抽出部
500
712
710
713
70
711
入力音声
701
71

【書類名】要約書
【要約】
【課題】　各音韻の認識率を向上することが可能な特徴パラメータの計算方法を用いた音声認識装置を提供する。
【解決手段】　音声認識装置１０において、特徴抽出部３０２は、入力音声信号３０１に対して、各々が所定時間長を有する時間窓に対応する複数のフレームを、順次増加する時間幅でスライドさせることにより、特徴パラメータの抽出を行う。単語辞書データベース３０４は、入力音声の音韻パターンにそれぞれ対応づけて標準パターンデータを格納する。認識処理部３０３は、特徴抽出部３０２で抽出された特徴パラメータと標準パターンデータを照合することで、対応する音韻を認識し、認識結果を出力する。
【選択図】　図３

特願２００３－２７７６６１

出　願　人　履　歴　情　報

識別番号　　［５０３１２１１０３］

１．変更年月日　　２００３年　４月　１日
［変更理由］　　新規登録
住　所　　東京都千代田区丸の内二丁目４番１号
氏　名　　株式会社ルネサステクノロジ